An independent verifier for machine-learned spacetimes, with a reproducibility audit of a neural black-hole search

Andrey Pluzhnik

Preprint, 12 August 2026. Not peer-reviewed.

Abstract. Neural networks are increasingly used to produce approximate solutions of the Einstein field equations, and their quality is normally reported using the same loss functions that were optimised to produce them. We present an independent evaluator that measures curvature from a metric callable by finite differences and shares no code with the systems it audits, calibrated so that its own error floor is published: it returns Ricci residuals of order 10^{-10} on analytic Schwarzschild, a Petrov speciality index of 1 to within 4×10^{-16} on a type-D metric and to 1.7×10^{-6} on a *trained network* imitating one, and it localises deliberately corrupted metrics. Applying it to five retrainings of the published black-hole configuration of the AInstein project (arXiv:2607.05489), which differ only in random seed, we find that the algebraic type and the trapped region reproduce in 5 of 5 runs while the vacuum condition holds in 2 of 5, with the independent residual spanning a factor 4.8 against a reported training-loss spread of 1.21. In this sample the reported loss ranks the runs in the opposite order to independently measured vacuum quality (Spearman −0.90, two-sided $p = 0.037$, $n = 5$, exploratory). We argue that a self-reported loss is insufficient evidence of geometric quality in this setting, and release the evaluator and all artifacts, including failed runs.

1. Why an external instrument

A physics-informed network that outputs a metric is trained to minimise a residual of the field equations, and the same residual is then quoted as evidence that the output solves them. This is not circular in principle — the residual is a real quantity — but it becomes fragile in practice for three reasons. The reported number is usually a sample average of a squared, weighted contraction, so it can be small while pointwise curvature errors are not; it is computed with the same differentiation kernel that shaped the optimisation, so shared bugs cancel; and it is evaluated on the sampling distribution used for training. None of these is a criticism of any particular paper. They are reasons why an outside measurement is worth making, in the ordinary sense in which an outside measurement is always worth making.

Our contribution is deliberately modest in scope: an evaluator that can be pointed at any system producing a metric, whose own error floor is measured and published, and a discipline that makes its verdicts checkable rather than assertable.

2. The evaluator

Given a callable returning the metric $g_{\mu\nu}$ at a point, we compute Christoffel symbols, the Riemann tensor, the Ricci tensor and the Kretschmann scalar by nested central differences in float64, choosing the step size by convergence sweep rather than by default. From the Weyl tensor we form an orthonormal Lorentzian frame, the complex Weyl operator $Q = E + iB$, and the algebraic invariants $I = \tfrac{1}{2}\sum\lambda_i^2$, $J = \det Q$, giving the speciality index $S = 27J^2/(4I^3)$, which equals 1 exactly for algebraically special geometries. For trapping we use the 2+2 split: the areal radius R read from the angular block, its gradient in the two-dimensional Lorentzian block, and $\Xi = h^{ab}\partial_a R\,\partial_b R$, together with the two null directions of the block; $\Xi<0$ with both null derivatives of R negative is a future-trapped surface. Where the coordinate time is spacelike the orientation is reported as ambiguous rather than guessed.

The evaluator imports no code from the audited system. Metric values cross the boundary as plain float64 arrays produced by a subprocess in an isolated environment, so a bug in the audited implementation cannot propagate into the measurement.

3. Calibration, stated before use

Check	Measured
Analytic Schwarzschild is vacuum (Schwarzschild coordinates)	8.7e−11
Kretschmann against $48M^2/r^6$	1.1e−9 rel.
Analytic Schwarzschild in the audited paper's Penrose and stereographic chart	6.7e−9 … 1.5e−8
Areal radius from the metric against an independent Lambert-W r(T,X)	3 decimals, 12 pts
Speciality index on a type-D metric	|S−1| = 4.4e−16
Speciality index on a *trained network* imitating Schwarzschild	|S−1| ≤ 1.7e−6
Trapping on the analytic metric	4/4 out, 8/8 in
Apparent horizon of a trained network imitating Schwarzschild	X = T ± 0.009
Deliberately corrupted metric	detected, localised
float32 storage of metric components	203× inflation
Finite-difference step, h ∈ [3e−4, 1e−2]	stable to 5 digits

The two network-based controls matter most. A trained network that approximates a type-D vacuum still reads as type D through this pipeline, and still places its horizon where the textbook one is, which is what licenses interpreting a departure as physics rather than as neural-approximation noise.

4. The audit

The AInstein project [1] reports neural metrics that are numerically Ricci-flat, algebraically general and possess trapped interiors. Its trained candidates are not published, so the object of audit is necessarily the published configuration rather than the reported candidates; this distinction is maintained throughout, and a request for the checkpoints has been sent to the authors.

We retrained the committed black-hole configuration five times, changing only the random seed, with one documented deviation: the committed `epochs: 10` debug value was set to 500. Each resulting metric was evaluated on 24 hidden points whose sampling seed is derived deterministically from the checkpoint's own SHA-256, so the evaluation coordinates cannot be chosen after seeing a result. Pass criteria were derived from a separately retrained neural Schwarzschild baseline, then committed and version-tagged **before** any candidate existed.

Seed	Reported final loss	Independent vacuum median	Vacuum	Speciality index	Trapped
126	1.32e−2	0.228	pass	2.29–2.56	11/12
124	1.18e−2	0.233	pass	2.27–2.55	11/12
125	1.10e−2	0.317	fail	2.30–2.58	11/12
123	1.11e−2	0.915	fail	2.29–2.58	11/12
127	1.09e−2	1.100	fail	2.29–2.58	11/12

Three observations follow, in decreasing order of confidence.

The reproducible parts and the fragile part are different parts. The speciality index and the trapped region appear in every run, and the index is reproducible to 0.43 % relative standard deviation across seeds. The vacuum condition — the field equation itself — holds in two runs of five, spanning a factor 4.8. The instability is specific to the Einstein term.

Two of five runs satisfy all three properties simultaneously. The configuration does produce candidates of the advertised kind, but not reliably, and the failure mode is not localised: the worst run's residual is elevated across the whole sampled Penrose block (exterior median 1.098, interior 0.986 over 95 grid points) rather than near the horizon or the diagram boundary.

The reported loss does not track independent quality in this sample. It varies by 1.21× where the independent residual varies by 4.8×, and the ordering is inverted: the lowest-loss run has the worst independent residual and the highest-loss run the best (Spearman −0.90, two-sided $p = 0.037$; Pearson −0.63, $p = 0.26$). We label this **exploratory**: the hypothesis was not pre-registered, it was noticed in the sweep output, and $n = 5$. We report it because it is inexpensive for others to test and because it is precisely the failure mode an external instrument exists to detect.

We also locate the apparent horizon per model by bisecting $\Xi = 0$ to ±0.003 in X. The analytic control returns the exact line X = T, a trained network imitating Schwarzschild returns it to within 0.009, and the candidates return near-vertical surfaces at X ≈ 0.63–0.71: a different horizon shape, not a displaced copy.

5. What this does and does not establish

All three measured properties are explicit targets of the audited objective, which includes an Einstein term, a speciality-index profile centred near 2, and a trapping term with weight 25 against the Einstein term's weight 1. Measuring them therefore corroborates that the published architecture achieves what it was designed to achieve, as judged from outside for the first time. It is not evidence that a new exact solution exists, and it is not a verdict on the authors' own candidates, which remain unpublished. A failed replication is not a refutation.

An incidental measurement is worth recording for anyone auditing such systems: the repository's supervised seed model, trained to match the analytic metric componentwise, has *worse* curvature than a physics-informed network at 6 % of training (Kretschmann error 1.5–8.2 % against 0.4–1.7 %). Pointwise metric agreement is not evidence of geometric agreement.

6. Limitations

No qualified relativist has reviewed this work. The sample is five seeds of one configuration. Hidden-point sets differ per seed by construction, which protects against coordinate selection but means runs are not compared at identical points. The marginally trapped surface is located only to bisection resolution, and the angular direction was sampled at a single stereographic point. The four-dimensional equivalent of the audited loss was not replicated exactly; the comparison on the authors' own scale is restricted to the two-dimensional case, where the square root of their reported loss and our pointwise maximum agree to 0.7 %. Training ran on CPU because consumer-GPU float64 measured ten times slower.

7. Availability

The evaluator, the workbench that computes stage statuses from hashed artifacts, all measurements and all failed runs are released under MIT. Stage statuses, thresholds and attestations are reproducible with

```
uv sync --all-groups && uv run sb verify-all ainstein-audit && uv run pytest -q
```

Validation thresholds were frozen and version-tagged before the results they judge existed; attestations record the git commit, the environment lock hash and per-artifact SHA-256 digests, and become invalid automatically when an artifact changes.

AI use. An AI agent (Claude, Anthropic) wrote the code and executed the experiments under the author's direction. The author set the research question, imposed the constraint that the agent may never mark its own scientific work as verified, required that discrepancies be resolved by measurement, and performed all external communication. Deterministic gates and preserved artifacts, not the agent's assertions, are what make the results checkable.

References

[1] E. Hirst, T. Schettini Gherardini and A. G. Stapleton, *Black Hole Black Boxes: Numerical Black Hole Metrics via AInstein Neural Networks*, arXiv:2607.05489 (2026). Code: github.com/xand-stapleton/ainstein, branch `blackhole`, commit 54736e46.